Consolidated Response

Questions from Innovation, Science, and Economic Development Canada (ISED)

Leadership Council on Digital Research Infrastructure (LCDRI)

Advanced Research Computing (ARC) Position Paper

Question #1: Estimates for Compute and Storage Growth 2

Question #2: Limitations to Private Sector Partnerships in ARC 3

Question #3: Instances of Canadian Researchers Choosing Not to Work in Canada 4

Question #4: What Gaps Would be Filled by the Local Layer Within a Strengthened Platform 4

Question #5: Regional Layer Budget Allocation 6

Question #6: ARC User Support: What Layer is Responsible? 7

Question #7: Regional vs. National Communities of Practice 8

Question #8: Option Differentiation by User 8

Question #9: Explanation of Current Target 9

Question #10: Equitable Distribution of Costs Under New Options 9

Question #11: Targeted Ranking for Processing Capacity Per Researcher 10

Question #12: Provincial/Regional Contributions to the National Layer 11

Question #13: Breakdown of Proposed Operating Costs for the National Layer 11

Question #14: Breakdown of Local Host Contributions 12

Question #15: Australia's Cost-Sharing Model 13

Question #16: Challenges Related to the Current Compute Canada Funding Model 13

Question #17: Pros and Cons of Proposed Funding Options 14

Question #18: Areas of Research Supported Through Various Level of Investment 14

Question #1: Estimates for Compute and Storage Growth

Would it be correct to say that these estimates for compute and storage growth are significant underestimations, not only because they do not consider projected user growth, but also because they do not take into account increased research activity as a result of new federal investments (e.g. Pan-Canadian AI Strategy)? (p. 17)

We believe that it is fair to state that our estimates likely represent a significant underestimation of potential future demand, particularly in the areas of computing and active storage. We are just at the cusp of implementing and understanding the full impact of new digital technologies. In the same way that twenty years ago we did not understand how the Internet would fundamentally reshape the ways in which we do business, communicate with one another, and solve problems in our daily lives (e.g. ten years ago, who would have predicted that 300 hours of video would be posted to YouTube every minute?), we do not yet fully appreciate how digital research infrastructure (DRI) will change the practice of research.

This makes projections for demand extremely challenging, as we cannot say with any certainty what researchers' needs will be as the number of researchers using ARC grow and they become more comfortable with it and knowledgeable on how to exploit its potential (via training such as that provided through investment by Compute Canada and its regional partners).

Therefore, the projections for demand for data analysis and active storage that we have made are based largely on history. They do, of course, incorporate user growth, but, as we state above, history can be an unreliable guide without the benefit of understanding fully the impact of critical unknowns such as how researchers will understand and use ARC in the future. This issue is compounded by the fact that, given that we have never met the full ARC requirements of Canada's research community, we are not working from a strong baseline in our understanding of current need. In addition, there are many other critical unknowns that will have a major effect on ARC usage and delivery such as: (a) emergent research areas (i.e. artificial intelligence (AI) and quantum computing), (b) major technology leaps (i.e. technological advance in new instruments such as next generation genomics sequencers), (c) changes in the research funding landscape (i.e. new funding in research areas that are heavily ARC-dependent); and (d) the exponential increase in the volume of data available from an ever-expanding range of new sources.

In addition, due to an insufficient supply of Canadian ARC resources, some researchers are accessing ARC resources outside of the country. This also contributes to the challenge in estimating future demand, as some of these individuals may choose to move their work back to Canada, if it becomes possible for them to do so.

All of this said, while the future estimates that we have provided are at best a guess, they do represent a level of investment we are confident that we can absorb and fully utilize.

Question #2: Limitations to **Private Sector Partnerships in ARC**

"Increasingly services and technology suitable for the needs of researchers are becoming more available "off the shelf," making private sector partnerships in ARC even more attractive, creating a fee-for-service model that is economically preferable for some uses." Can you clarify how this is an issue/challenge? (p. 21)

There is debate within Canada's DRI community about the current ability of Canada's private sector to offer the type of specialized computation and storage that is needed by researchers and to provide more cost-effective service than what is being delivered by Compute Canada and its partners. It is possible to imagine a time when at least some of the computation and storage needs of Canada's research community could be delivered in partnership with the private sector, however, we are not there yet. Commercial offerings are designed to meet the needs of the business user - advanced computing requirements are, by definition, "advanced" and are not currently well-served by profit-driven private sector cloud providers.

Unfortunately, should the potential for increased partnership with the private sector materialize, the current ARC funding model would make it difficult for Canada's ARC community to take advantage of this opportunity. The current ARC funding model works much better for capital acquisitions than for fee-for-service models. In areas where fee-for-service would be more economical, the current funding model isn't agile enough to allow us to take advantage of cost savings. In addition, the external technology environment is evolving much more rapidly than our funding cycles allow us to adapt.

Regardless of whether or not it may eventually be possible to partner with the private sector on fee-for-service models, this partnership would likely provide only infrastructure and software. It is unlikely that it could or would provide the advanced technical support services that researchers need to access and optimize computing and storage resources. This could cause significant challenges and disruption to services for researchers if not managed and planned for appropriately. For example, it is difficult to provide technical support without the benefit of direct access to the infrastructure platforms or of having experience with them. Any decision to move to a partnership model with the private sector would need to be grounded in a delivery model that ensures strong integration and coordination between those providing the infrastructure and those providing the support services to researchers.

Question #3: Instances of Canadian Researchers Choosing Not to Work in Canada

"Anecdotally, we know that some researchers have chosen not to work in Canada because they could not access the ARC resources that they needed for their research programs." Can the LCDRI provide good examples of this? (p. 21)

While we can't point to a flood of researchers who are leaving Canada because of a lack of ARC resources, many universities have highlighted the challenge of recruiting new faculty who require this access. There are many factors that contribute to the decision to accept a faculty position in Canada and ARC is one of them. In our report we described this observation as being "anecdotal"; however, concrete examples do exist, especially in the field of bioinformatics. Due to privacy issues, we are not able to access and provide personal information on researchers who have chosen not to work in Canada due to a lack of ARC resources. We are, however, able to provide the following illustrative examples

- UBC missed an opportunity to recruit a chemist from Harvard, who at the time needed about 12 million compute hours for his research program; and,
- A genomics researcher from the Genome Institute of Singapore refused a job offer from the Université de Montréal due to the lack of sufficient computing resources.

Question #4: What Gaps would be Filled by the Local Layer Within a Strengthened Platform

"[The local layer] also supports practical implementation of operational details and fills in gaps when ARC resources are unavailable through the regional/provincial and national layers." What gaps would there be if the ARC platform is strengthened? How does the local layer fill these gaps? By purchasing their own systems? (p. 25)

A healthy ARC platform is both diverse and agile enough to ensure that the full scope of researcher requirements is met and that full operating efficiencies are achieved. There are a number of cases in which local ARC systems would continue to be the most efficient and effective infrastructure and service delivery layer for researchers.

First, a national system is not a practical or efficient layer for meeting the needs of a research team that has lab/observation equipment that is streaming data on a second-by-second basis. It is very difficult technically for a national layer to provide the connectivity to accommodate the vast amount of data that would be generated by this instrumentation. A local ARC solution is required in these circumstances.

Second, there are a number of research areas that require secure and/or separate access to the ARC and the network. An example of this would-be research that is incompatible with remote or shared system environments such as cyber-security labs through which researchers are running experiments on the dark web or setting "honey pots" to study and attract computer viruses in the wild. Given the nature of this work, it would be a significant security risk to run these experiments through a national system. They must be undertaken through local network and ARC access. There are also circumstances in which certain research projects require researchers to be hands-on in the frequent reconfiguration of infrastructure.

Third, like regional systems, local systems provide an important space for experimentation and innovation in way that larger, less agile, national systems can't. This can lead to the development of innovative tools that may be essential for a small group of researchers, but not for the system as a whole and, most importantly, to the opportunity to explore new technologies, at minimal risk to the national layer, that may be candidates for scaled delivery by the national layer in the future. For example, the IceCube Neutrino researchers at the University of Alberta purchased some experimental GPUs through a CFI grant that they are willing to share with other researchers across Canada when they are not using them. WestGrid is coordinating the sharing of these resources within the broader Canadian ARC community. In short, regional and institutional experimentation add a critical layer of resiliency to Canada's ARC platform that should not be ignored.

Fourth, regional and local layers often provide critical last mile connectivity, ensuring that there is an appropriate network connection between research lab, or data collection point, and the main university connection. Often, while an institution itself may provide adequate connectivity and ARC centres are very well connected to each other, there is a bottleneck for researchers in connecting their research lab or data collection point to institutions and ARC centres because this last element of connectivity is missing. Further, firewalls can limit the data transfer capacity from labs. Therefore, special care has to be taken to organize data transfer, such as the use of Science DMZ[1], and this is a service that is best organized locally to support access to national infrastructure.

Lastly, it would be difficult for the national layer to provide the kind of basic, day-to-day services that researchers need. While the national layer could support the delivery of highly specialized ARC services, the local layer, in partnership with the regional layer, would deliver a more accessible and general level of service in an environment that is much more conducive to encouraging greater and more effective ARC usage. This is because front-line service providers are closer to researchers and therefore better able to tailor their services to the needs of their local communities and because researchers are often most comfortable working with front-line service providers who are located in close geographic proximity to them, such as on campus, and with whom they have an existing relationship.

[1] A Science DMZ offers a network environment that is tailored to meet the needs of high performance science applications, including large data transfers, remote experiment control, and data visualization. It addresses common network performance problems for participating research institutions.

In summary, a national layer is essential for ensuring the greatest possible access to ARC for Canadian researchers, but large systems cannot replace all requirements for local infrastructure and service delivery, nor are they able to provide the agility and responsiveness needed to generate the innovation that will keep Canada's ARC platform resilient for the future.

It is important to stress that we are not suggesting a return to an ARC centre at each institution or in every lab across Canada. But, we do believe that there is a balance required - something between consolidating all ARC in Canada in one site with a mandate to provide infrastructure and services that meet every conceivable requirement of every researcher and research program across Canada, and every lab at every university having their own ARC infrastructure and services.

Question #5: Regional Layer Budget Allocation

On what exactly would the $20M/yr for the regional layer be spent (operations of regional consortia)? (p. 27)

The estimated $20M/year for the regional layer would be spent on the following operational activities:

- piloting new innovative services regionally/locally in response to user demand and in an environment that is more agile than that provided by a larger national system (this would ensure that ARC in Canada would strike an important balance between having a large, stable, national system and retaining its ability to be innovative and push boundaries through an environment of continuous learning and experimentation);
- addressing regional/provincial ARC-related research priorities;
- providing support services to researchers and front-line service providers;
- delivering regional/provincial user training to researchers and front-line service providers, with an emphasis on new user engagement and training;
- developing HQP at all points in the pipeline to keep up with technological changes etc. (i.e. students and professional ARC staff);
- coordinating regional/provincial communities of practice;
- providing strategic advice and support to the national layer;
- user need and satisfaction assessment;
- coordinating key initiatives and relationships with regional/provincial governments and other funders, as well as institutions and other DRI partners (i.e. NRENS);
- awareness-raising and ARC strategic policy development; and,
- supporting specialty services e.g. high availability/fail-over clusters, high security, PHIPA compliance and providing an environment in which smaller local initiatives can be reasonably and helpfully consolidated.

It is important to remember that we are proposing that the total federal investment in the regional layer would be $10M - $14M, depending on the option that is selected.

Question #6: ARC User Support: What Layer is Responsible?

Isn't providing the primary level of ARC user support for academic researchers the responsibility of the local layer rather than the regional layer? (p. 27)

Providing ARC support is a specialized skill and it should not be confused with the provision of IT services, which is a very different practice. In large institutions, such as the University of Toronto or UBC, it may make sense to have dedicated ARC support services as there would be enough of a critical mass of users in these institutions to use full-time support. But, this would not be the case for smaller institutions. Not only would they find it difficult to fund full-time staff in this area, but it would be inefficient and ineffective to do so, as it is unlikely that they would have enough researchers requiring access to these specialized services on a regular basis (it should be noted that it would also be a significant burden for larger institutions to manage the increased costs associated with ARC support). Pooling resources at the regional level allows everyone to benefit from important economies of scale, providing the critical mass required to ensure that ARC services exist, as well as ensuring a more equitable level of access to high quality support services for all researchers, regardless of their location or the size of their institution.

Pooling resources at the regional layer also allows support teams to exchange information and to continuously update their skills more easily. For instance, when you pool support across a region, instead of just locally, and use a ticketing system, users and support staff learn with every ticket to which a response is provided. This leads to greater efficiencies and an improved user experience, as the responses and advice are consistent and developed with the advantage of being able to draw on the skills and experience of a wider group of experts.

Similarly, sharing ARC support services at the regional, and in some cases national level, makes important economic sense in highly specialized areas such as genomics where the costs of having a full-time expert on staff in most, if not all, institutions across Canada would be prohibitive and inefficient as it is unlikely that this resource would be needed by any one institution on a full-time basis. Leveraging and sharing expertise within a broader regional or national context just makes sense. For instance, the genomics hotspot for WestGrid is in BC, due to the Genome Sciences Centre and UBC Medical school. WestGrid is able to leverage BC's expertise regionally so that they don't have to have bioinformaticians at every site locally. This approach achieves an important fine balance between having the support that is needed close to the users and in which they can feel confidence, while also providing a model that is cost effective and efficient.

Question #7: Regional vs. National Communities vs. National Communities of Practice

For what would regional communities of practice be responsible, and how would they differ from national communities of practice? (p. 27)

The Legal Dictionary defines a community of practice as: "A network of peers with shared skills in a profession. Each member is driven to help each other". As such, they are, as their name suggests, community-driven and most often organically developed. While dedicated funding and human resources can increase their success, their structures are not generally overly formalized. Most importantly, communities of practice are, at their root, learning communities that enable individuals who share a common interest, responsibility, or profession to leverage each other's knowledge and skills, as well as that of others who may be invited to contribute to the group.

Given the member-driven and organic nature of communities of practice, it is not possible for us to say, at this point, exactly what communities of practice would need to be supported at either the regional or national layers. However, we can say with certainty that communities of practice are important and extremely helpful to building strong and resilient DRI that is able to meet the needs of researchers across Canada, regardless of the size of their institution or where they are located. We can also say that to be most successful, we need to ensure that communities of practice are supported with maximum flexibility in mind, allowing researchers, or those who serve them, to define their own needs and communities (e.g. in Digital Humanities, French speaking researchers tend to collaborate more with colleagues in France, Switzerland or Belgium than with English speaking colleagues in Canada).

Lastly, we need to facilitate groups that allow for diversity of interest, as well as scope of membership. For instance, in some cases, a community of practice may develop around a group of individuals working at a local institution on a particular project. In other cases, communities of practice may coalesce regionally or nationally around groups of people who are working in similar fields across institutions (they can also have multiple layers, with representatives from the local layer sitting with others at the regional or national layer).

Question #8: Proposed Option Differentiation by User

"Build on the same foundations that are articulated in Option 1 to ensure that the needs of small and moderate users are met, and that Compute Canada's current target of meeting 75% of large user needs is achieved;" Don't options 3 and 4 meet 100% of large user needs? (p. 29)

When developing and describing the various options, we were careful to avoid the suggestion that any of them would serve 100% of user need, as there will always be some researchers who can design problems that match the available computing capacity.

However, we did distinguish among four categories of researchers: small, medium, large, and largest users. All of the options are designed to meet fully small and medium user needs. Options 1 and 2 would meet 75% of large user need and little or none of the needs of our largest users. You are correct that Options 1+2 and 2+3 would meet all of the large user need and all but a few of the largest users. We apologize for the confusion. It is not clear in our document.

Question #9: Explanation of Current Target

Why is 75% the current target? How was this target determined, and why does it have merit? (p. 29)

This target corresponds to the current Compute Canada Key Performance Indicator in this area. This target recognizes the fact that the top 2% of researchers use half of the computing cycles currently provided by Compute Canada and its regional partners.

While not meeting the needs of all researchers in Canada, the 75% target that is suggested in Options 1 and 2 would help to ensure that researchers would continue to need to optimize their code and efficiently use resources, while also increasing overall access to ARC resources. In addition, while it would not meet fully the needs of Canada's largest users, this target, modest as it is, would represent a major improvement over the current state of affairs for large users.

Question #10: Equitable Distribution of Costs Under New Options

"Ensure that these costs are distributed equitably, so that individual universities are not using funds from their institutional budgets to provide services to other universities across the country." How do the three options ensure this? And under what circumstances would an institution use their funds to provide services to other universities under this model (if operational funding is provided through the national layer)? (p.29)

Currently, the CFI funding model requires a 40/40/20 (federal/provincial/other) sharing of costs. As a result, institutions, which generally provide the 20% match, that choose to compete for the right to host a national site, are assuming 20% of the costs associated with running that system on behalf of other institutions across Canada. This issue is amplified in jurisdictions where the 40% provincial match for operations or capital is not provided, creating a situation where some institutions that compete for the right to host a national site are assuming 60% of the costs of running the site on behalf of other institutions across the country.

We are fast approaching a time when, as the ARC needs of Canadian researchers grow and the percentage of the budget that goes into operating expenses grows from about $10M a year in 2018 to $27M a year in 2023, the current competitive model will become

unsustainable and unscalable for the institutions that are using their own institutional budgets to pay 20% - 60% of the costs of running a national ARC system.

The new model that we are proposing would eliminate this unsustainable distribution of costs, by removing the requirement for those institutions that are managing a national ARC system to provide 20% - 60% of the costs. In our new model, costs would either be assumed fully by the federal government in Options 1 and 1+3, or by a federal/provincial cost-sharing arrangement in Options 2 and 2+3.

It is very important to note that none of this suggests that institutions would stop making significant contributions to ARC. Our proposed options just provide greater clarity in the system and ensure that funding responsibilities are better aligned with the layer at which responsibility for delivery has been established.

Question #11: Targeted Ranking for Processing Capacity Per Researcher

"Bring Canadian ARC capabilities to a target of #6 or #8 in processing capacity (gigaflops) per researcher as compared to other countries, making us (roughly comparable to Italy or France)." Why did this change from previous draft material (used to be a higher ranking)? (p. 28)

We presume that the ranking that was provided in a previous draft was an error. This ranking is the correct one.

However, it is important to note that these numbers have the potential to be fairly fluid. Some of the countries in the top 15 in gigaflops per researcher processing capacity are quite small in terms of their total number of researchers (e.g. Switzerland and Saudi Arabia). Therefore, these countries can raise their ranking dramatically with a single large investment. For nations with larger research communities, the only way to stabilize their ranking is through consistent investment, such as has been the case in the US and, more recently, in China. As stated in our position paper, we believe that increased ARC funding that is both predictable and sustainable, as proposed in Options 1 and 2, would stabilize Canada within a ranking of #6 - #8 in processing capacity (gigaflops) per researcher, recognizing that setting the goal as a range is more reasonable given the fluidity of these rankings. Ultimately, we would like to be at or near the top of this ranking, but for now a more modest positioning still represents a major step forward for science in Canada.

Question #12: Provincial/Regional Contributions to the National Layer

"Ensure an equitable distribution of provincial/regional contribution to the federal layer" How will an equitable distribution of contributions among regions/provinces be ensured under this option? (p. 30)

As stated in the answer above, the current CFI model requires a matching contribution from provinces for the capital and operating costs associated with hosting a national site. As a result, those provinces that have competed for the right to host a national site, are agreeing to contribute matching funds for the costs associated with running that system on behalf of other provinces across the country. For example, only two provinces (and two regions) provided matching funding for the capital purchases through the CFI Cyberinfrastructure Fund, Stages 1 & 2.

The options that we are proposing would remove this issue, as in Options 1 and 1+3, the federal government would provide full funding for the national layer, negating the need for provincial investment, and Options 2 and 2+3 would require a formal federal/provincial cost-sharing agreement. In addition, the options would allow provincial governments and regions to invest in their own environments against priorities and needs.

Question #13: Breakdown of Operating Costs for the National Layer

Re: $29M/yr for operating the national layer: Why are operating costs so much higher than current levels of operational funding, particularly if funding for user support personnel is to be provided through the regional and local layers? Is it possible to provide a breakdown of operating costs? (p. 31)

The breakdown for "steady state" operating costs for the national layer, as projected for the 2022/2023 fiscal year, include the following:

Item	**FY23**
Power	$12.6M
Maintenance. & Repair	$3.0M
Personnel	$13.4M
Total	$29.0M

We are proposing a consistent average capital investment of $60M per year, which is much higher than in previous years. This leads to increased operational costs over current levels, as a result of the following:

- doubling of the implied power costs over 5 years, even taking into account the saving due to the retirement of older, less energy efficient systems;
- tripling of maintenance costs over 5 years; and,
- doubling of the staff involved system operations; this also includes 6 FTE for cybersecurity, an area in which we have been currently deemed to be inadequately resourced.

An inflation rate of 2% per year has also been taken into account.

Question #14: Breakdown of Local Host Contributions

What is the share of the ~$50M or ~$75M that is from host institutions? How was this estimate developed? In the LCDRI's proposed model, do non-hosting institutions contribute to the total costs of the ARC platform (including capital)? If yes, is that contribution included in this estimate? (p. 32)

The lower estimate of $50M was derived by summing several individual conservative estimates of both capital and operations costs. First, we estimated conservatively vendor in-kind contributions based on a $60M capital investment for a total of $24M. We then added a conservative capital investment by institutions of $3M (eg. data centre improvements such as cooling system and power capacity upgrades). In terms of an estimate on contributions from host institutions for operating costs, we estimated $23M (i.e. $9M personnel, $10M in power, $2M in maintenance and repairs, and $1M in traditionally unrecognized institutionally borne costs of security, cleaning, plant maintenance, and commercial insurance costs). Of the conservative $50M estimate, the total contribution absorbed by institutions is, therefore, $26M.

On the projection of $75M, we estimated vendor in-kind contributions at 75% of capital investment, for $45M in vendor contributions, and institutional capital investments (eg. data centre improvements) at $5M (Note: this is roughly the amount absorbed by institutions in Stage 1 of CFI Cyberinfrastructure funding). In terms of host contributions to operating expenses, we estimated a contribution of $25M (i.e. $9M personnel, $12M power, $3M maintenance and repairs, and $1M in traditionally unrecognized institutionally borne costs of security, cleaning, plant maintenance, and commercial insurance costs). Of the $75M estimate, the total contribution from institutions is, therefore, $30M.

Question #15: Australia's Cost-Sharing Model

How is Australia able to maintain a 33/66 cost-sharing ratio between federal and state governments? Do they face challenges that are not outlined in this description? (p. 44)

The Australian cost-sharing ratio is apparently not highly prescriptive, but rather collaborative, and does not always follow the exact ratio noted. In terms of challenges, although there is a national supercomputing strategy (as part of the national research infrastructure strategy), there is no national supercomputing "organization" -- instead there are several supercomputing providers (providing cycles as well as services/expertise), all of which are collaborative, and each has to cobble together the required co-investment. This funding complexity has been noted as a problem by the Australian research community. The governance of these collaborative ventures is also complex, and the challenges of coordinating the different ventures have also been raised as an issue.

Question #16: Challenges Related to the Current Compute Canada Funding Model

How exactly does the current model for funding Compute Canada hamper its ability to strengthen the ARC platform through national services (e.g. develop shared software, implement cyber-security protocols)?

At present, regions/provinces and institutions are paying 60% of the costs associated with funding ARC in Canada. Each of these regions/provinces and institutions has their own board, committees and/or Cabinets who have their own terms and conditions, as well as requirements and priorities, for how the funding that they provide is spent. This can make it very challenging to implement national services that align with everyone's priorities.

Moreover, as we explain in questions #10 and #12, the current CFI funding model requires a 40/40/20 (federal/provincial/other) sharing of costs. As a result, provinces and institutions that compete on the right to host a national site assume 60% of the costs associated with running that system on behalf of other institutions across Canada. We are fast approaching a time when, as the ARC needs of Canadian researchers grow and the percentage of the budget that goes into operating expenses grows from about $10M a year in 2018 to $27M a year in 2023, the current competitive model will become unsustainable and un-scalable for the institutions that are using their own institutional budgets to cover the costs of running a national ARC system and more difficult for provinces to justify within their own provincial budgets.

Lastly, the lack of predictability in ARC funding in Canada has also significantly affected its ability to undertake critical long-term system and national services planning that is

essential to ensuring better service to researchers and to achieving greater system efficiencies to optimize investment.

Question #17: Pros and Cons of Proposed Funding Options

The first two funding options present different federal investment levels ($98M/yr and $79M/yr) that, based on the descriptions given, achieve the same outcomes. What are the pros and cons of each option?

We understood that it was important to provide options to you for the Minister's consideration, so we developed two funding investment options for the national and regional layers that accomplish the same outcome - one through which the federal government would pay the full costs of the national layer and 33 % of the regional layer for a total of $98M/yr and one through which the federal government would contribute 75% of the costs of the national layer and 50% of the costs for the regional layer for a total of $79M/yr and the regional later would pay 50% of the costs for the national layer for a total of $10M/yr.

With the first investment model, the federal government would need to consider how to ensure that the regional voice was represented at strategic planning and other critical tables at which collaboration and coordination are essential. With the second model, the federal government would need to negotiate and put in place a mechanism for cost-sharing with provincial governments.

Both options would have the benefit of ensuring a more sustainable funding model compared to what we have today. With Option 2, one could argue that it's beneficial to leverage a diversity of funding sources to alleviate the federal burden. However, adding more funders in the second option, would also require more funding priorities to continue to align and this would create a more complex and potentially challenging governance and management structure.

Question #18: Areas of Research Supported Through Various Level of Investment

Can you elaborate on what areas of research will be adequately supported through the various levels of investments proposed? Which disciplines, domain communities, or specific "large users", will particularly benefit through these investments?

The two figures below provide a view of the CPU usage (Figure 1) and the number of active groups per discipline in Compute Canada (Figure 2) between April 1, 2016 and March 31, 2017. As viewed in Figure 1, the disciplines with the highest CPU usage are chemistry/biochemistry, physics, and engineering. Figure 2 shows that engineering, biological/life sciences, chemistry/biochemistry, mathematics/computer science, and physics have the greatest number of active groups, with engineering and biological/life sciences having the most.

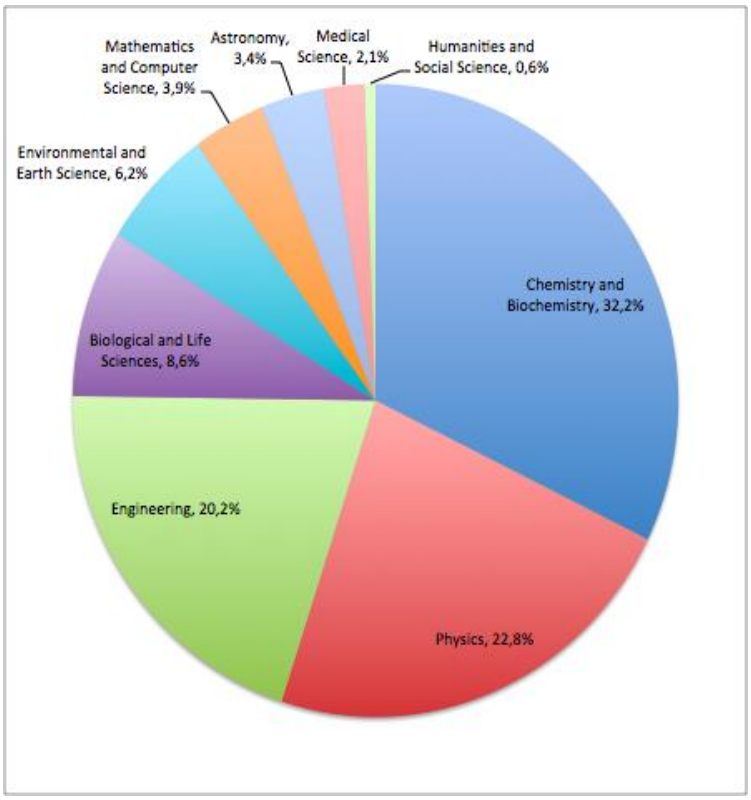

Figure 1: CPU usage per discipline in Compute Canada (in percentage) between April 1, 2016 and March 31, 2017

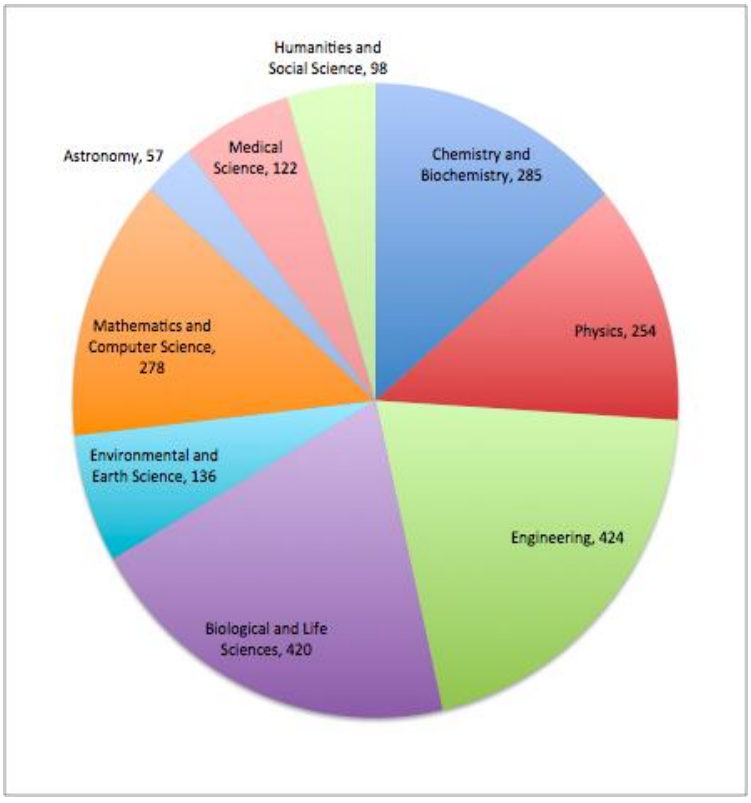

Figure 2: Number of active groups per discipline in Compute Canada (here an active research group has launched at least one job) between April 1, 2016 and March 31, 2017.

In terms of active storage, currently, the communities with the largest needs include: genomics, astrophysics, and particle physics. In addition, when images are used, such as for the self-driving car, AI and Big Data also become large users of active storage, as are large international collaborations and platforms such as those for brain imaging and metagenomics. We can presume that these discipline groups will continue to be our highest users of computing and storage into the future and that they would benefit the most from new federal investment in ARC. However, as new user communities develop, become more comfortable with the technology, and more creative in their use of it, these ratios may change. And, additional investment will be essential to augmenting the services that are critical to building this increased user adoption and research output. While many academic disciplines and researchers have recognized the potential of ARC and have embraced its use, many others have not, and widely accessible user services are needed to encourage and support researchers in their use of ARC.

It is also important to remember that large, targeted, federal investment in new areas such as AI could also skew these ratios. Lastly, it should be noted that, given the way in which the options for new funding are proposed, only 75% of our large users and none of our largest users in the areas listed above (i.e. engineering, biological/life sciences, chemistry/biochemistry, mathematics/computer science, and physics), would have their needs fully met with Options 1 and 2. Examples of researchers and research teams who are our large and largest users include climate and earth sciences, astronomy and cosmology, material science, and computational chemistry.

The Canadian ARC system must serve a diverse set of needs. In particular, it must serve a very large number of researchers with modest computational needs at the same time it serves a smaller group of researchers with very large computational needs. This can be a difficult balance to achieve with the resources that we currently have. Researchers who must perform single, very large calculations require large computational capacity in a single system. While that single system can also be shared by many researchers with modest needs, it is designed such that the entire system could be allocated to a single calculation for some period of time.

The largest such system in Canada at the moment is Niagara, currently being installed by Compute Canada at the University of Toronto. As such, our largest users will be limited to single calculations that use about 60,000 cores at the same time. There are single machines in other countries with millions of cores (up to nearly 20M), allowing researchers in those countries to tackle much more complex problems. Generally speaking, this limitation affects those modeling complex systems such as climate, nuclear physics in stars, interactions of molecules, and combustion in engines. Canadian researchers in these areas currently look abroad to find machines capable of modeling their most complex systems. Researchers in astrophysics, such as Ue-Li Pen (University of Toronto, Director of CITA), Petr Nvratil (TRIUMF), and Falk Herwig (University of Victoria) have RAC allocations on Compute Canada systems but are very much limited by the size of the largest machines Compute Canada has to offer. Similarly, researchers such as Peter Tieleman (University of Calgary, T1 CRC chair in

Biomolecular Simulation), Regis Pomes (Hospital for Sick Children Research Institute, both working in the field of drug design) and Andre Bandrauk (Université de Sherbrooke, a pioneer of atto-second physics) are limited in their research by the scale of the largest Canadian systems. David Zingg's (University of Toronto, Institute for Aerospace Studies) studies of computational aerodynamics or Clinton Groth's (University of Toronto, Institute for Aerospace studies) studies of combustion are limited in a similar way. Canada would need to build a significantly bigger single machine (e.g., at least 5x larger than Niagara) to effectively serve Canadian researchers who need to perform very large modeling studies and to build research capacity in these areas by attracting world-class researchers from abroad. Operating at this scale would require a significant increase in investments such as that suggested in Options 1+3 and 2+3.